农业图书情报学报
DOI：10.13998/j.cnki.issn1002-1248.23-0849

人工智能科学家对科学认识论的挑战

段伟文[1,2,3]
（1. 中国社会科学院大学 哲学院，北京 102445；2. 中国社会科学院哲学所，北京 100732；
3. 上海人工智能实验室，上海 200232）

摘　要：[目的 / 意义]本研究旨在探讨人工智能科学家可能给科学认识论带来的挑战。[方法 / 过程]科学发现一直是人工智能研究的重要主题，人工智能科学发现的下一步是发展人工智能科学家，即能进行自主和自动化科学发现的人工智能系统，其研究质量与最优秀的人类科学家的水准无法区分。回顾人工智能在科学研究中的相关应用之后，阐述了人工智能科学家最为重要的特征及其研究计划的核心，在此基础上提出了人工智能科学家在认识论层面带来两个根本性的改变：人工智能能力跃升和人工智能驱动的科学研究范式嬗变。[结果 / 结论]对于相关科学认识论问题的讨论需要走出一般的哲学式论辩，面向即将到来的人工智能科学革命提出了搁置否定性的批评、关注过渡期的难题、动态追踪可能的突破口、寻求更好的类比等 4 个认识论策略。

关键词：科学智能（第五范式）；人工智能科学家；科学认识论；科学自动化
中图分类号：TP18；B017；N02　　文献标识码：A　　文章编号：1002-1248（2023）11-0004-09

引用本文：段伟文. 人工智能科学家对科学认识论的挑战[J]. 农业图书情报学报, 2023, 35(11)：4-12.

最近，作为非人类的人工智能大模型 ChatGPT 入选《自然》年度十大人物，堪称科学发展的里程碑式事件。由此不难联想到足球机器人世界杯（RoboCup）创始人、日本系统生物学家北野弘明（Hiroaki Kitano）受 RoboCup 的启发，曾于 2016 年撰文指出，要对人工智能提出一个新的重大挑战，开发一种能够做出重大科学发现、乃至可获得诺贝尔奖的人工智能系统[1]。2020 年初，英国阿兰·图灵人工智能研究所召开了主题为“向人工智能科学家提出重大挑战：人工智能系统能够实现诺贝尔级的发现”的研讨会。北野弘明与倡导科学中“深思熟虑的人工智能”的南加州大学计算机学家尤兰达·吉尔（Yolanda Gil）[2]和倡导“科学自动化”的罗斯·金（Ross King）[3]共同担任会议的主席，与会者发出了推进诺贝尔图灵挑战赛的全球（行星）倡议。次年，北野弘明在《自然》杂志旗下的《系统生物学与应用》撰文进一步提出，应通过发起诺贝尔图灵挑战赛，发展出能进行顶级科学研究的高度自主的人工智能系统，它们有望做出重大科学发现并在 2050 年获得诺贝尔奖。时隔两年之后的 2023 年 10 月，北野弘明来到颁奖时节的斯德哥尔摩，主持了一场主题为如

收稿日期：2023-10-17
基金项目：国家社会科学基金重大项目“智能革命与人类深度科技化前景的哲学研究”（17ZDA028）；中国社会科学院哲学研究所创新工程项目“前沿科技的哲学基础与科技时代的价值重置”（2024ZXSCX06）
作者简介：段伟文（1968- ），中国社会科学院大学哲学院教授，博士生导师，中国社会科学院哲学研究所研究员，E-mail：duanweiwencass@163.com

何发挥人工智能在科学研究中的创造性作用的研讨会，进一步考虑为取得世界级科学成果的人工智能以及人工智能与人类合作设立奖项。

正如北野弘明所指出，科学发现一直是人工智能研究的重要主题。从早期的 DENDRAL 与 META-DENDRAL，到后来的 MYCIN、BEACON、AM 和 EURISKO，人工智能科学发现走过了近 80 年的探索之路。最近，生物学等领域中出现的自动化实验系统，已经可以实现假设生成、实验规划和执行的闭环。在其看来，虽然这些自动化实验系统局限于单个数据集或运用有限资源的特定任务，但却清晰地昭示出人工智能用于科学发现的下一步是开发一个能够自动进行科学发现的系统，并以此具有颠覆性科学研究方式带来重大发现。而加速这一历史进程的最佳方法莫过于通过明确的使命宣言制定大胆而具有挑衅性的目标——让人工智能科学家通过图灵测试——其研究质量与最优秀的人类科学家的水准无法区分，甚至可能获得诺贝尔奖之类的象征性成果，最终创造出人类科学发现活动的替代形式[4]。

1 人工智能科学家及“科学的科学”的认识论意义

人工智能在科学研究中的应用由来以久，其传统领域属于机器发现或发现信息学等科学发现的认知研究，可以追溯至西蒙（Herbert Simon）20 世纪 40 年代的研究和始于 20 世纪 60 年代的 Dendral 等第一代人工智能专家系统。后者由斯坦福大学的人工智能学家费根鲍姆（Edward Feigenbaum）和诺贝尔生理学或医学奖获得者、遗传学家莱德伯格（Joshua Lederberg）等合作开发，旨在帮助有机化学家通过质谱分析和利用现有化学知识识别未知的有机分子。该系统两个主要项目之一的 Meta-Dendral 是作为维京号火星探测器的一部分而设计的，其初衷是在人类科学家无法肉身抵达的遥远的火星构建一个自动化的科学研究系统。这一研究无疑是跨学科合作的典范，参与者中还包括机器学习先驱布坎南（Bruce G. Buchanan）和著名化学家杰拉西（Carl Djerassi）。在此被称为“发现科学”的发展脉络中，另一个备受关注的是由西蒙推动的将归纳推理形式化的 BEACON 系统，其目的是发现经验规律，一般会借助数据驱动的启发式方法。该系统声称重新发现了开普勒行星运动定律、欧姆定律、斯涅耳定律、动量守恒、万有引力和布莱克比热定律等科学定律。还有研究者沿着类似的进路，在没有任何物理、运动学或几何方面的先验知识的情况下，发现了哈密顿量、拉格朗日量以及其他几何和动量守恒定律[5]。但争议之处在于，由于该系统及类似方法是在清理过的干净数据之上拟合出科学方程的，与开普勒等科学定律的真实发现过程有很大不同。尽管如此，BEACON 依然是人工智能科学发现的重要里程碑。

近年来，基于人工神经网络的深度学习等大大推进了科学研究自动化步伐。这一进路可追溯至 20 世纪 80 年代末，“神经网络”一词刚开始激发公众想象力时，粒子物理学家意识到机器学习正好可用于在复杂粒子探测器无数相似的读数中寻找微妙的空间模式。他们花了很多时间才消除了人们对机器学习等可能是魔法、骗术与黑箱的疑虑，如今人工智能技术已成为在高能物理和宇宙探测器的数据洪流中探索发现的物理学家的标准工具之一[6]。进入 21 世纪，沿着计算机辅助发现、数据驱动的科学和 AI 驱动的科学发现的方向，自动化科学发现在生物和化学等领域蓬勃发展。对此，CONNOR 等的综述《化学科学中的自动发现第一部分：进展》[7]和《化学科学中的自动发现第二部分：展望》[8]作出了系统介绍。从自动发现的视角出发，他们对发现进行了再定义——按照发现的内容将化学中的发现区分为物理物质（分子、材料、设备）、过程和模型 3 种类型，探讨了如何将它们统一为搜索问题。他们进而将自动发现归纳为基础计算推理、框架机理模型的发现、化学过程的非迭代发现与迭代发现以及新物理物质的非迭代发现与迭代发现等进路，探讨了合成化学、药物发现、无机化学和材料科学领域由计算机辅助和自动化加速或生成发现的案例。由此不难看到，随着计算机、数据挖掘、机器学习等数字技术和人工智能的迅猛发展，科学实验和建模的性

DOI：10.13998/j.cnki.issn1002-1248.23-0849

质正在发生前所未有的改变。值得指出的是，针对如何评估一项发现在多大程度上可以归因于自动或自主发现，该综述对所需要考虑的问题提出了建议。在他们看来，尽管提及自动或自主发现往往关注发现的“闭环”程度——其中隐含地预设了多个“假设 - 测试 - 修正信念”循环的迭代过程——进而暗含了实现完全的科学自动化的可能性，而目前大多数自动发现的案例最好被描述为不能没有人类干预的“开环”。

但将完全的科学自动化作为人工智能研究终极大挑战的人工智能科学家的倡导者则有更大的志向，他们不愿接受人工智能是科学家越来越好用的工具这种渐进式的前景，而主张人工智能科学家最为重要的特征恰在于对人工智能作为科学研究工具的突破和超越。对此，2020 年阿兰·图灵人工智能科学家会议的与会者明确指出：“我们一致认为，人工智能科学家和人工智能科学工具之间存在根本区别”。在他们看来，人类科学家通常使用各种人工智能工具来完成有助于发现的特定任务（例如数据分析、文本提取等），但这些人工智能工具不能被视为人工智能科学家，因为它们只解决科学过程的狭隘方面。更重要的是，它们在设定目标、解释结果和交流发现方面缺乏像人类科学家那样的自主性。因此，诺贝尔图灵挑战赛明确的目标是，开发具有高度自主研究能力的人工智能系统，而非开发人工智能工具[9]。

对这一激进愿景，北野弘明强调人工智能科学家研究计划的核心应是“科学的科学”，而非“人类科学家的科学过程的科学”[4]。促使他提出这一观点的理由是人类在科学认知上的局限性。他指出，在生物医学研究中，存在一些超出人类认知能力的根本性困难，而随着系统生物学的出现，这个问题变得更加明显。他认为，人类在科学中的认知局限主要涉及信息处理、知识表达和认知偏差等方面，包括：①信息视界（Information Horizon）问题，如生物医学研究有太多的数据和出版物，其产生速度远远超出了人类的信息处理能力；②信息鸿沟（Information Gap）问题，即科学论文所用的语言经常存在歧义、不准确和信息缺失的情况；③表型不准确问题，表型即对生物异常的表征和分类，因其往往基于研究者的主观解释和共识而难免很不准确；④认知偏差问题，研究者在推理和交流中会不可避免地使用模糊的自然语言和符号，其思维过程不可避免地存在偏见；⑤少数报告问题，即与大多数报告的共识不同的少数报告是否可以当作错误或虚假报告而被丢弃，如何区分其中的错误报告和那些可能促进重大发现的报告[1]。

正是由于人类科学家在科学研究中存在越来越大的认知局限性，人工智能科学家的倡导者认为，科学必将突破人类主导的范式，将来的科学应该由可能打破人的认知局限的人工智能科学家主导。据此，将有一种不同于人类主导科学过程的新科学，或者说人工智能科学家主导的未来科学将成为更一般意义上的科学，故以走向这种更一般的科学为目标的人工智能科学家计划堪称“科学的科学”。因此，发起人工智能科学家诺贝尔挑战赛将在认识论层面带来两个根本性的改变：一方面，科学家不再绝对意味着人类科学家，科学研究中的人工智能将实现从认知自动化工具到自主的自动化认知主体（智能体）的飞跃；另一方面，科学活动和科学认识不再必然是人类的科学活动或科学认识。这意味着科学研究和发现将不必然是人类主导并基于人类经验感知和推理的人的认知活动，而可能走向人工智能主导并基于智能体经验感知和推理的一般智能体的认知活动。

在科学哲学和科学技术研究等领域，对北野弘明等所论及的“科学的科学”即关于科学研究活动规律的研究，一般称之为元科学或科学学——也可以视为广义的科学认识论。以往旨在更好地理解科学是如何运行的科学学的对象默认为人类的科学活动过程，而人工智能科学家之类的自动化科学发现系统的构建本身实则是对自动化科学研究这一全新的科学形态的创建。因此，恰如费曼所言“我无法创造的东西，我就无法理解”，自主的人工智能科学家在科学学上的挑战显而易见：通过创造另一种形式的科学发现加速科学研究的进程，使科学展现其更一般性的本质，更好地造福我们的文明。

2 科学自动化愿景：费根鲍姆测试与人工智能能力挑战

不论称其为机器发现、发现的科学、还是科学的科学，机器人科学家或人工智能科学家属于人工智能探索与具体的科学发现相互交叉的跨学科研究，其共同愿景的最大公约数是不断提升科学自动化。从人工智能探索和科学发现这两个视角来看，人工智能科学家扮演着不同的角色：当我们从人工智能的角度研究科学自动化时，用于科学发现的人工智能和机器人系统是实验对象，其实验领域是人工智能和机器人学；而当我们使用人工智能来自动发现科学领域的新知识时，人工智能和机器人系统就是进行科学研究和实验的主体[10]。因此，当我们谈到科学自动化所带来的革命性变化的宏大愿景时，实际上涉及对人工智能的科学发现能力跃升和人工智能驱动的科学研究范式嬗变两个方面。本节先讨论前一个方面。

从人工智能能力跃升的视角看，如果依然将人工智能科学家的能力视为对人类科学家能力的模仿，人工智能科学家所面对的认知能力挑战可以简单地表达为图灵测试（TT）的人工智能科学家版本——如作出诺贝尔奖科学发现的人工智能科学家是否进化到与顶级人类科学家难分伯仲？但问题是，就人工智能目前的发展水平而言，通过日常生活意义上的图灵测试尚不多见，谈何通过诺贝尔奖图灵测试呢？面对这一现实，被誉为专家系统之父并获得 1994 年图灵奖的费根鲍姆曾提出费根鲍姆测试（FT）。

在 2003 年发表的《计算智能的一些挑战与重大挑战》一文中，费根鲍姆将其测试建立在两个基点之上。其一，强调人的智能和人工智能可以归结为计算智能："当科学家使用'智能'或'智慧'这些术语时，他们指的是人类认知行为的大集合——人的思维。当生命科学家谈到动物的智能时，他们要求我们回忆起一系列人类行为，并断言动物能够（或不能）做到这些。当计算机科学家谈到人工智能、机器智能、智能体（代理）或计算智能时，我们也是在指一系列人类行为。虽然智能意味着人的思考，但我们也许可以用计算来复制同样的行为。事实上，现代认知心理学的一个分支就是基于这样一个模型：人类的心智和大脑是复杂的计算'引擎'，也就是说，我们自己就是计算智能的典范[11]。"

其二，明确指出为了能够在复杂的智力任务上表现出高水平的性能，甚至可能超越人类的水平，人工智能（计算智能）必须拥有该领域的广泛知识。他接受了图灵奖得主计算机学家吉姆·格雷（Jim Gray）建议的图灵测试的方案：在图灵测试中，人工智能至少获得 30%的胜率。格雷主张的测试内容既包括图灵最初提出的模仿游戏还加上了"像人类一样阅读和理解、像人类一样善于思考和书写"[12]。费根鲍姆指出，这意味着对人工智能的智能考察必然涉及人类活动的广度、深度和关注点，人工智能所拥有的计算智能需要一个庞大的知识库。但鉴于获取如此庞大的计算机可用知识库是一项非常艰巨的挑战，费根鲍姆为此提出了图灵测试的替代方案。其基本思想是聚焦于对具有某个学科知识的人工智能体的推理质量（复杂性、深度）的测试，并强调为了避免他的修订方案"污染"图灵的思想遗产，建议将此称为费根鲍姆测试。

费根鲍姆测试的具体做法是，在每次测试中，由一位来自国家科学院的精英科学家担任评议者，同时对一个某个科学领域具有科学推理能力的人工智能体和一名来自国家科学院自然科学、工程或医学领域的人类精英科学家提出专门的科学问题——如要求解释某一科学理论或天体物理现象，看人类评委能否以高于概率的准确率评判出哪个是他在国家学院的同事，哪个是电脑？参照格雷的计算智能成功标准，如果人工智能体在 3 场学科评判竞赛中"赢得"一场，即如果 3 位评委中有一位无法在人类和人工智能体之间做出可靠的选择，就可以认为挑战成功。

费根鲍姆指出，根据当时专家系统的发展水平，这是一项艰巨的大挑战。在他当时看来，一旦出现超智能计算机，会导致的悖论是超智能计算机很容易与人类精英区分开来，因为它将具备优于人类的归纳和解决问题的能力，提出严谨、可信、正确甚至有趣的

DOI：10.13998/j.cnki.issn1002-1248.23-0849

解决方案和理论。也就是说，超智能计算机的出现可能意味着费根鲍姆测试的失效。但真正有意义的问题是如何向超智能计算机跃迁？

对此，他与人工智能社群的主流认知是，计算智能的发展瓶颈是知识库。人们早在此前 20 多年就意识到人工智能领域的科学缺失，但因此而兴起的机器学习却事与愿违。虽然最初人们认为机器学习将学习由符号实体、关系和本体建立起来的符号概念，却很快从符号概念走向了统计学的边界，实体和关系在本质上均被当作统计学的。尽管这对当时方兴未艾的数据挖掘起到了推波助澜的作用，但对作为计算智能关键和核心的大型知识库基本上没有影响。为了应对人工智能依然存在的科学缺失，他认为计算智能下一步应直面两大挑战：先是在文本阅读的基础上构建大型知识库，将知识工程的成本降低一个数量级；进而从万维网中提炼出一个庞大的知识库，将知识工程的成本降低几个数量级。

尽管受当时主流观念影响，费根鲍姆未能预见到基于联结主义的深度学习和生成式人工智能的成功，但他却预见了当前大模型知识生成的奥秘所在。他不无洞见地指出：“可以把万维网视为世界上最大的数据库，尤其是如果把所有通过链接获取的信息都包括在内的话。它包含了世界上的许多时事、文化历史文本和其他信息类型的知识积累，堪称我们人类知识的一面镜子，或许更准确地说，它是知识的一次转型。对于试图构建能通过图灵测试或费根鲍姆测试的计算智能的知识工程师来说，它无异于知识树上诱人的苹果[11]。”

沿着与费根鲍姆为人工智能设置重大挑战以促进其发展类似的思路，尤兰达·吉尔、罗斯·金和北野弘明等诺贝尔人工智能挑战赛的倡导者提出了一系列挑战目标。他们指出，实现诺贝尔图灵挑战赛需要几十年的艰苦研究，为此可以定义中间的重大挑战以驱动各种智能能力的发展：①展示和形式化有关某些科学领域的专家级知识。②由人工智能系统驱动、用于科学发现的实验机器人。③出版传播。生成、总结、讨论、评论、批评 / 比较科学文章，提出令人信服的科学问题和 / 或研究计划来回答问题。他们还强调，此类挑战应针对特定的科学学科进行定义，这将促进不同科学领域为应对挑战展开协同合作。例如，在生物医学和生物技术领域的诺贝尔级发现可能成为人工智能科学家的目标，包括发现生物学的新原理、逆转衰老、治愈癌症、了解大脑如何产生行为、高效的药物发现和疫苗生产、创造具有特定功能的生物体等；而诺贝尔图灵挑战赛可能面临的重大挑战可能包括：由医疗 / 社会动机驱动生物医学研究、开发高度自主和集成的研究平台、现有生物医学存储库的生物信息学分析等[9]。在 2023 年发表的《评估人工智能驱动的科学自动化的框架》一文中，罗斯·金等进一步指出，要在图灵挑战赛中取得成功需要克服巨大的技术挑战，人工智能系统需要具备以下能力：①对其研究目标做出战略选择；②在超越限制的领域产生令人兴奋和新颖的假设；③设计新颖的方案和实验来测试超出使用原型实验的假设；④以人类科学家可以理解的方式关注并描述重大发现[13]。

3 人工智能驱动的科学研究范式嬗变及其科学认识论挑战

再看人工智能的科学范式嬗变，不论将其概括为 AI for Science，还是“第五范式”都可以看到由此所带来的科学研究范式的全新变革。像所有的新科学研究进路的倡导者一样，当前机器人科学家、人工智能科学家的倡导者乐于为我们描述一个笃定完美的科学未来。在他们看来：“科学的未来在于人工智能主导的闭环自动化系统。它们在科学研究的全周期自主运行，从假设生成到实验验证和结果重新解释不断迭代。这些系统将模仿人类的科学过程，但工作速度更快、更精确。它们将减少偏见，并能够开辟更大的科学发现领域[14]。”但不论如何，在人工智能驱动科学研究方兴未艾的今天，我们都应该对由此必然导致的科学研究范式的嬗变及其探索实践中的认识论挑战作出系统分析和前瞻性考量。

毋庸置疑，构建能进行自主和自动化科学发现的

人工智能系统是十分巨大的挑战，但其倡导者对此抱持乐观主义的态度，并在科学实践层面不断推进相关研究。另一位重要的倡导者罗斯·金是这方面研究的先驱，他将自动化的机器人实验系统称为“机器人科学家”。所谓机器人科学家即能在物理空间实现的实验室自动化系统，它可以利用人工智能技术来执行相对完整的科学实验周期，从中可见人工智能科学家的雏形。在一些特定的实验自动化系统中，机器人科学家已经可以自动提出假设和解释观察结果，设计实验来测试这些假设，使用实验室机器人物理运行实验、解释结果，然后重复这个循环。

罗斯·金等开发的第一代机器人科学家“亚当”（Adam）自主生成了有关酿酒酵母的功能基因组学假设，并通过实验室自动化对这些假设进行了实验测试。他们研制的第二代机器人科学家“夏娃”（Eve）特别针对被忽视的热带疾病的药物开发，将计量经济学模型用于药物筛选，使得药物发现在经济上优于标准药物筛选[15]。而开发中的第三代机器人科学家“创始者”（Genesis）旨在并行自动化数千个闭环实验周期，自动学习真核细胞的计算模型。其研发者认为，这一模型是现代科学中最重要和最具挑战性的任务之一——构建细胞的高保真模型将需要数千个循环的设计实验和模型改进，而当前的系统生物学研究很少完成单个循环的模型改进。

严格地讲，人工智能科学家依然在路上。目前的机器人科学家依然只是某种形式的人工智能驱动的自动化实验仪器自动运行系统，当今最好的人工智能系统无法自行定义其假设空间和实验设计，充其量可以算作人工智能科学家的初级形式。显然，人工智能驱动科学研究或走向人工智能科学家的具体路径必然受到人工智能前沿发展的影响。最近，生成式人工智能和大模型成为风口，卡内基梅隆大学化学工程系的博伊科（Daniil A. Boiko）等在《自然》杂志撰文，提出一种被称为“联合科学家”（Coscientist）的人工智能体，它可以根据简单的人类提示来规划、协调和实施化学研究周期中的多项任务。例如，当被要求合成特定分子时，Coscientist 通过互联网搜索以设计合成路线；它能为所需的反应设计实验方案；编写代码来指引移液机器人；还可以从反应的结果中学习，提出修改方案，通过迭代使反应过程不断优化[16]。

透过机器人科学家和联合科学家等人工智能科学家的初级形态，有助于探寻从作为科学研究工具的人工智能走向人工智能科学家的演化过程。显然，这种演化的前景不应该简化为像人类科学家的人工智能科学家实体的涌现，而应该将其纳入人工智能驱动的科学研究范式的嬗变中，深入探讨其在科学认识论层面的引发的颠覆性挑战。而这一挑战与科学实在论、科学划界等科学认识论问题的根本差异在于科学认识的主体不再必然是人类。不无微妙的是，这一不易避免的趋势是由人类借助人工智能实现的认知能力增强所带来的悖论。对此，科学哲学和计算哲学家保罗·汉弗莱斯（Paul Humphreys）尖锐地指出了这一无法掩盖的事实：作为“万物的尺度”，在我们对自己的感知能力进行扩展的同时，科学认识论将不再是人类认识论[17]。其具体表现为以下两个方面。

一是从人工智能体的认知能力来看，人工智能与其他科学装置结合，共同构成了人类科学家的认知增强工具，但其发展是否意味着人类正被逐出科学的中心、科学终将走向独立于人类的科学。汉弗莱斯指出，包括人工智能在内的计算机模拟和计算科学使得人类的观测、推理等认知得以延伸，人类科学认知的边界不断拓展，令人类自然认知能力的限制得以超越，但问题是科学会不会成为独立于人类的认知活动，或者说是否能够完全免于人类参与呢[17]？

二是从人工智能体在科学活动的泛智能体协同网络中的地位来看，人工智能体或人工智能系统与人类科学家群体智能的组合正在构成科学研究和发现的泛智能体网络，先进的人工智能系统将置于这一网络的中心，巧妙而有效地协调的这种“半人马”式的智能协同与人机共生形式。但由此会导致的问题是，人类被逐出科学认知乃至决策系统中心之后，必然会带来科学的意义和价值何在之类的新科技人文两难。对此，北野弘明预言，未来应用于科学发现的先进智能将超越人工智能和人类专家的现有组合；但也指出，这条

DOI：10.13998/j.cnki.issn1002-1248.23-0849

道路最终是否会通过促进一系列重大科学发现让我们的文明更加强大，拟或由于对人工智能系统的广泛和过度依赖而更加脆弱，尚待观察[1]。

4 余论：面向人工智能科学革命的若干认识论策略

虽然这种独立于人类的科学所带来的冲击是革命性的，但人工智能驱动科学研究的最新探索却越来越有力地支持其实现的可能性。在此态势下，对于相关科学认识论问题的讨论不能不走出一般的哲学式论辩，转而探寻面向即将到来的人工智能科学革命的认识论策略。

（1）搁置否定性的批评。适当搁置有关人工智能科学认知能不能发现真正的科学问题和进行创新之类的抽象的否定性论述，更多地关注人工智能科学家的建构实践过程中的相关问题及其可能的解决途径。不再过多纠缠于机器人科学家不能绝对自主且人始终应在回路之中、人工智能只能做常规科学而无法推动科学革命之类的囿于概念的讨论，转向关注向人工智能科学革命过渡的探索实践中的具体问题。例如，面对日益复杂的科学研究，人类的认知局限、错误、偏见等缺陷将显得更加严重，智能系统如何帮助人类科学家克服其局限性，避免人的认知偏见影响科学发现，更好地应对科学研究不可重复性之类日益严重的隐忧。

（2）关注过渡期的难题。推动人工智能科学革命的关键是设法消除过渡期可能导致新的范式退却和失败的瓶颈。这些问题包括人工智能认识不透明问题和人工智能科学研究的鲁棒性问题等。对于前者，其出路可能在于承认机器智能与人类智能在机制上的平行性，转而通过机器智能之间的比较揭示其认识不透明性。对于后者，应关注人工智能科学研究和发现中对研究数据与过程干扰等展开的对抗研究，探讨将人工智能科学研究的稳定性、安全性作为其评价指标的认识论意义。

（3）动态追踪可能的突破口。科学革命和创新源于新颖的组合，人工智能科学革命的实现需要从认识论上进行优化整合设计。由此，可以运用互补策略和渐进策略。例如，当今的自动化智能系统大多是为执行特定任务而构建的，知识面非常狭窄，而人类可以通过巧妙和创造性的视角、非传统的洞察力、科学问题和目标的优先排序以及对成果和想法重要性的认识等与之形成互补。随着人工智能科学研究系统的发展，可以为其设置人工智能工具、人工智能助手、人工智能合作者等渐进的阶段性目标。

（4）寻求更好的类比。在人工智能发展过程中，人们常常会借助类比认识其发展中出现的各种问题。对于机器人科学家和人工智能科学家，为了揭示其与人类科学家可能的关系，可以引入亚人类、准人类、超人类等新的类比，促使我们思考其发展的各种可能性。例如，针对生成式人工智能研究中广为关注的“幻觉”说，南加州大学计算机科学家尤兰达·吉尔最近表示，这实际上表明大型语言模型非常适合在科学研究中进行头脑风暴。而另一位研究者指出，语言模型可能会产生不准确的信息并将其呈现为真实的信息，但这种“幻觉”实际上意味着某些看起来真实的东西，而这正是假设[18]。

参考文献：

[1] KITANO H. Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery[J]. AI magazine, 2016, 37(1): 39-49.

[2] GIL Y. Thoughtful artificial intelligence: Forginganewpartnershipfor-data science and scientific discovery[J]. Data science, 2017, 1(1/2): 119-129.

[3] KING R D, ROWLAND J, OLIVER S G, et al. The automation of science[J]. Science, 2009, 324(5923): 85-89.

[4] KITANO H. Nobel Turing Challenge: Creating the engine for scientific discovery[J]. NPJ systems biology and applications, 2021, 7: 29.

[5] SCHMIDT M, LIPSON H. Distilling free-form natural laws from experimental data[J]. Science, 2009, 324(5923): 81-85.

[6] CHO A. AI in Action: AI's early proving ground: The hunt for new particles[J]. Science, 2017, 357(6346): 20.

[7] COLEY C W, EYKE N S, JENSEN K F. Autonomous discovery in

the chemical sciences part I: Progress[J]. Angew. Chem., Int. Ed., 2020, 59: 22858–2893.

[8] COLEY C W, EYKE N S, JENSEN K F. Autonomous discovery in the chemical sciences part II: Outlook[J]. Angew. Chem., Int. Ed., 2020, 59: 23414–23436.

[9] AI scientist grand challenge: Summary of discussion during workshop held in February 2020[EB/OL]. [2023-08-05]. https://www.turing.ac.uk/sites/default/files/2021-02/summary_of_discussion_workshop_2020_ai_scientist_grand_challenge_clean.pdf.

[10] SOLDATOVA L N, CLARE A, SPARKES A, et al. An ontology for a robot scientist[J]. Bioinformatics (Oxford, England), 2006, 22(14): e464–e471.

[11] FEIGENBAUM E A. Some challenges and grand challenges for computational intelligence[J]. Journal of the ACM, 2003, 50(1): 32–40.

[12] GRAY J. What next? A dozen information-technology research goals[J]. Journal of the ACM, 50(1): 41–57.

[13] OECD. Artificial intelligence in science: Challenges, opportunities and the future of research[M]. Paris: OECD Publishing, 2023.

[14] KING R, ZENIL H. A framework for evaluating the AI-driven automation of science, in Artificial intelligence in science: Challenges, opportunities and the future of research[M]. Paris: OECD Publishing, 2023.

[15] KING R D, SCHULER COSTA V, MELLINGWOOD C, et al. Automating sciences: Philosophical and social dimensions[J]. IEEE technology and society magazine, 2018, 37(1): 40–46.

[16] BOIKO D A, MACKNIGHT R, KLINE B, et al. Autonomous chemical research with large language models[J]. Nature, 2023, 624(7992): 570–578.

[17] (美)保罗·汉弗莱斯. 苏湛, 董春雨, 孙卫民, 校译. 延长的万物之尺: 计算科学、经验主义与科学方法[M]. 北京: 人民出版社, 2017.
HUMPHREYS P W. Extending ourselves: Computational science, empiricism, and scientific method[M]. Beijing: People's Publishing House, 2017.

[18] MATTHEW H. Hypotheses devised by AI could find "blind spots" in research[EB/OL]. [2023-11-17]. https://www.nature.com/articles/d41586-023-03596-0.

The Challenge of Artificial Intelligence Scientists to the Epistemology of Science

DUAN Weiwen[1,2,3]

(1. School of Philosophy, University of Chinese Academy of Social Sciences, Beijing 102445; 2. Institute of Philosophy, Chinese Academy of Social Sciences, Beijing 100732; 3. Shanghai Laboratory for Artificial Intelligence, Shanghai 200232)

Abstract: [Purpose/Significance] This study aims to explore the challenges that artificially intelligent (AI) scientists may bring to scientific epistemology. [Method/Process] Scientific discovery has long been of interest to AI researchers. The next big step in AI is the development of AI scientists. AI scientists should be able to independently motivate, make, understand, and communicate discoveries. Although the current robot scientists are still just a form of AI-driven automated experimental apparatus, and the best AI systems today cannot define their own hypothesis space and experimental design. At best, they can be considered to be a primitive form of AI scientists. Clearly, the specific path of AI-driven scientific research or the transition to AI scientists will inevitably be influenced by the

DOI：10.13998/j.cnki.issn1002-1248.23-0849

frontier development of AI. Current AI systems must overcome the following major technical challenges: 1) making strategic choices in their research goals; 2) developing the ability to generate exciting and novel hypotheses in areas that push boundaries; 3) designing innovative approaches and experiments to test hypotheses that go beyond the use of prototype experiments; 4) focusing on and describing important discoveries in a way that can be understood by human scientists. The highly autonomous AI scientists can either make discoveries on their own or collaborate with other human and machine scientists to make Nobel-level discoveries. After reviewing the relevant AI applications in scientific research, this study illustrates the main characteristics of AI scientists and the two disruptive changes they bring about at the epistemological level: a leap in AI capabilities and AI for Science as the 5th paradigm of scientific research. [Results/Conclusions] The implications of AI for Science are revolutionary, but recent AI-driven explorations in scientific research increasingly support the possibility of its realization. In this situation, discussions on the epistemological issues of relevant sciences need to go beyond general philosophical debates and instead explore epistemological strategies for the coming scientific revolution in AI. In view of the coming scientific revolution in AI, this study proposes four strategies. First, we should pay more attention to the problems and solutions in the process of developing AI scientists. Second, the key to advancing the scientific revolution in AI is to find ways to eliminate factors that may lead to failure. Then, we use different strategies to achieve the scientific revolution of AI. Finally, we take advantage of metaphorical methods to help us develop AI scientists.

Keywords: AI for Science; AI scientists; scientific epistemology; automated science